Anirudh Tunoori

Professor Imieliński

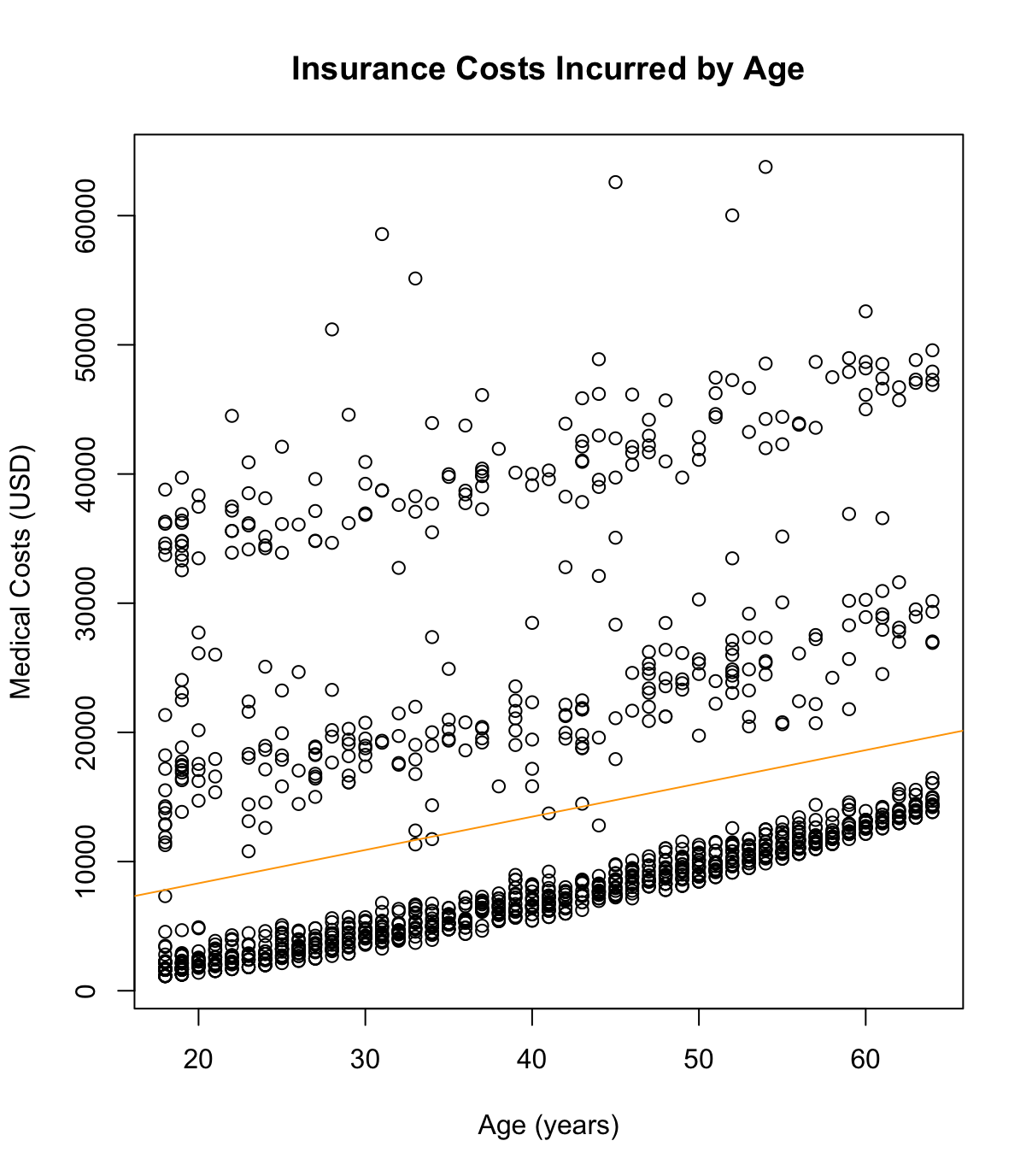
CS142: Data 101

Spring 2018

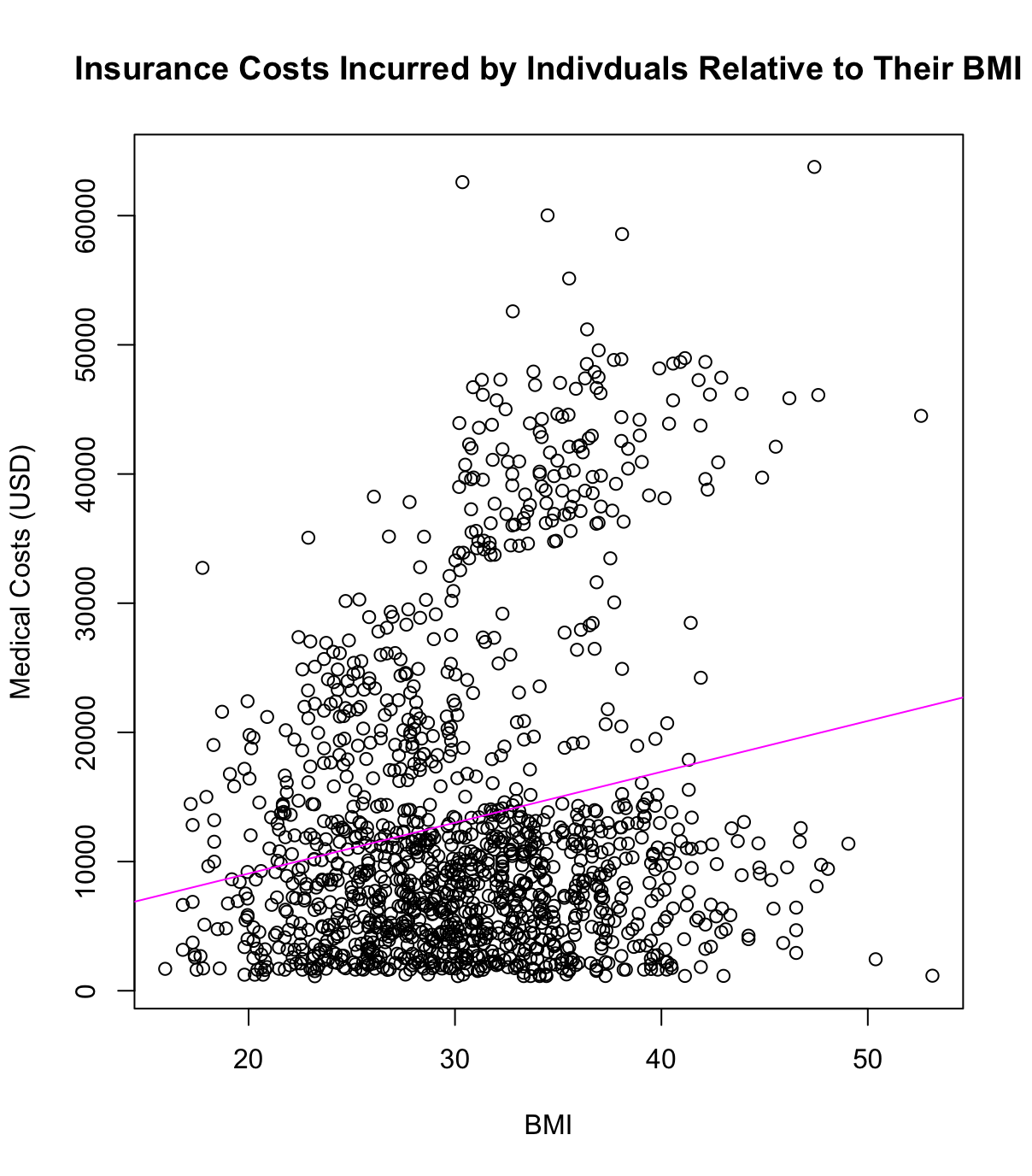
Forecasting Medical Insurance Costs

Welcome readers! I am Anirudh Tunoori. I am a junior majoring in Computer Science and Mathematics and minoring in Economics. I am taking a data literacy course: Data 101, and one of the things that I have come to appreciate is that though we may not always be aware of it, we are surrounded by data. Data plays a vital role in the modern world, its given rise to terms like “Big Data” and “Data Science”; which have become ubiquitous of late. Data has grown exponentially more powerful and the tools at our disposal to make sense of it and do useful things with it have become numerous. As a result, we are able to make important decisions, recognize previously enigmatic patterns, and formulate incredibly powerful questions/problems using data. Recently, while exploring different data sets, I came across a dataset that sought to predict medical costs of people based on specific personal attributes. I was interested in this dataset because health care and insurance/medical costs are a prevalent concern and exploring relevant data may help us in reforming existing systems. Namely, can we make reliable approximations of an individual’s insurance costs using a few variables? And if a prediction-based model exists, how can different groups make use of this information to improve the existing, underlying system? The data I looked at can be found at: <https://www.kaggle.com/mirichoi0218/insurance>.

First, I took a look at the relationship between the age and the charges incurred by the individual. Here is a scatterplot that depicts this relationship:



We can see within this scatterplot, three clusters, each depicting a positive correlation: increasing medical costs with increasing age. But the existence of clusters indicates that age is not the sole factor in the amount that an individual is likely to spend. The three clusters fall within the following monetary ranges: $0-$12,000, $12,000-$25,000, $30,000-$45,000. Within each cluster, the data points are relatively close and with increasing age individuals in each cluster can expect to be charged more as well. Finally, the regression line does not appear to be useful in the context of this plot.  
First, I took a look at the relationship between the BMI (Body Mass Index) and the charges incurred by individuals. I feel that this is an important relationship since medical experts often relate the BMI as a measure of an individual’s overall health as it (ratio of a person’s height to their weight) gives them an understanding of an individual’s physiological composition and because it enables them to classify people into different weight classes. Furthermore it enables the experts to form professional opinions on what individuals have to be wary of. Here is a scatterplot that depicts this relationship:



In this plot, the regression line indicates a slight positive correlation between the BMI and insurance costs. Yet, this is by no means statistically significant evidence. There are dense clusters within this plot likely as a result of the fact that significant proportions of the sample size fall within the same “body classification” (20-40). There are also several outliers, several of the data points depict high BMI’s and low medical costs and vice versa. Along with a few other reasons, this is likely a result of the fact that the incidence of certain diseases (certain cancers) have nothing to do with BMI.   
I wanted to examine the relationship between gender and insurance costs; perhaps certain treatments, medications, and procedures for one gender are costlier than for the other, thus resulting in higher insurance charges. To do so I decided to conduct a Z-test to determine if there is statistically significant evidence contained within this data set that suggests that one gender can expect to be charged more than their counterpart. When I did so, I obtained a Z-score of **2.10088** and a p-value of **0.0 1782541**, which is statistically significant evidence to conclude that males typically spend more on medical costs than females do as far as this sample is concerned. Though it should be noted that more males were surveyed in this dataset than females.